BLIND IMAGE QUALITY ASSESSMENT FOR MULTIPLY DISTORTED IMAGES VIA CONVOLUTIONAL NEURAL NETWORKS

Jie Fu[1,2], *Hanli Wang*[1,2,*], *Lingxuan Zuo*[1,2]

[1]Department of Computer Science and Technology, Tongji University, Shanghai 201804, P. R. China
[2]Key Laboratory of Embedded System and Service Computing, Ministry of Education, Tongji University, Shanghai 200092, P. R. China

ABSTRACT

The past decade has witnessed a growing development of Image Quality Assessment (IQA) techniques. However, the researches of IQA with multiple distortion types are still limited especially on blind image quality assessment methods. In this paper, a Convolutional Neural Network (CNN) based method is proposed to predict the quality of multiply distorted images without references. Inspired by the early human visual model, the proposed CNN based method combines feature learning and regression for estimating the quality of multiply distorted images. The proposed network consists of one convolutional layer, one pooling layer with max and average pooling, two full connection layers and one softmax classification layer. With this network structure, the relationship between the accuracy of CNN and the prediction monotonicity of IQA is explored. Experimental results on the newly released LIVE multiply distorted image quality database verify the effectiveness of the proposed CNN based method.

Index Terms— Blind image quality assessment, convolutional neural network, multiply distorted image, accuracy, prediction monotonicity.

1. INTRODUCTION

With the rising trend towards image quality requirements, Image Quality Assessment (IQA) [1] becomes an important topic for both the scientific research and application development of digital image processing systems. The goal of IQA is to build a computational model to evaluate image perceptual quality accurately and automatically [2]. Owing to the importance of IQA, it has been used in a wide range of computer vision and image processing applications, such as image processing and transmission systems [3], image/video compression [4, 5], restoration [6], etc.

*Corresponding author (H. Wang, E-mail: hanliwang@tongji.edu.cn). This work was supported in part by the National Natural Science Foundation of China under Grant 61472281, the "Shu Guang" project of Shanghai Municipal Education Commission and Shanghai Education Development Foundation under Grant 12SG23, and the Program for Professor of Special Appointment (Eastern Scholar) at Shanghai Institutions of Higher Learning (No. GZ2015005).

In the research arena of IQA, the Mean-Squared Error (MSE) and Peak Signal-to-Noise Ratio (PSNR) have prevailed for decades as the most popular IQA metrics before gradually giving way to the Structural Similarity index (SSIM) [7], which is proposed a decade ago. MSE and PSNR calculate the Euclidean-style distortion and they do not perform well with subjective fidelity scores. In order to improve the IQA ability, the SSIM metric extracts the structural information through a framework of measurement, comparison, as well as a combination of luminance and contrast, thus achieves better IQA performances than PSNR.

In general, three categories of IQA approaches can be classified, including Full Reference (FR) [7, 8, 9, 10], Reduced Reference (RR) [11] and No Reference (NR) (also known as blind) [12, 13, 14, 15] depending on the accessibility of reference images. As far as FR IQA is concerned, the reference image is given as a perfect version of the image and FR metrics aim at presenting computerized algorithms to evaluate the perceptual quality of each distorted image. As compared with FR, RR metrics use partial information of the reference image and NR metrics do not employ any reference image. Despite many IQA methods behave remarkably well for singly distorted images, it is still of challenge to exploit IQA approaches for multiply distorted images. In practice, images are usually contaminated by multiple distortion types such as noise, blur, compression, and so on. Therefore, the research efforts of this work are focused on blind IQA research for multiply distorted images.

On the other hand, deep neural network has recently gained researchers' attentions and achieved great successes on various computer vision tasks. Unsurprisingly, the Convolutional Neural Network (CNN) model which is one of the most representative deep neural networks is also applied to improve IQA performances. In [16], CNN is introduced into IQA research for singly distorted images with the network structure consisting of one convolutional layer with max and min pooling, two fully connected layers and one output layer. As compared with a number of state-of-the-art approaches, the IQA performances are greatly improved by this CNN based approach [16].

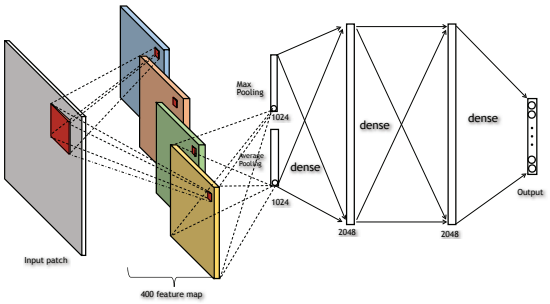

Fig. 1. Overview of the proposed CNN based method for multiply distorted image quality assessment.

Inspired by [16], a CNN based method is designed in this work for multiply distorted images. As compared with the CNN model employed in [16], a mixture of max pooling and average pooling is employed by the pooling layer. Moreover, the CNN model parameters are further regulated and optimized to fit for the IQA problem about multiply distorted images. To verify the performance of the proposed CNN based method, the recently released LIVE Multiply Distorted image quality database (LIVEMD) [17] is employed for experiments, with the comparative results demonstrating the superiority of the proposed approach over other state-of-the-art IQA approaches. The rest of this paper is organized as follows. The proposed CNN based method for multiply distorted images is detailed in Section 2. Section 3 presents the comparative experimental results. Finally, Section 4 concludes this work.

2. PROPOSED CNN BASED METHOD FOR MULTIPLY DISTORTED IMAGE QUALITY ASSESSMENT

2.1. Overview

The proposed CNN based method for multiply distorted image quality assessment is illustrated in Fig. 1, where it can be observed that there are six layers, including one input layer, one convolutional layer, one pooling layer, two full connection layers and one softmax classification layer.

Specially, the detailed architecture of the proposed CNN based method is designed as ① 32×32 ② $26 \times 26 \times 400$ ③ 2×1024 ④ 2048 ⑤ 2048 ⑥ N for the corresponding six layers, where N is the number of labels used in the last softmax classification layer. As inspired by the CNN structure in [16], the input layer employs 32×32 image patches which are locally normalized. Then, the convolutional layer filters input image patches with 400 kernels each of which applies a 7×7 filter with the stride equal to 1 pixel. As a result, the convolutional layer produces 400 feature maps each of which obtains the size of 26×26. After that, the pooling layer reduces each feature map to one max value and one average value followed by two fully connected layers of 2048 nodes each. At last, a linear regression with an N-dimension output is performed to generate the final IQA estimation.

2.2. Convolution

In the convolutional layer, the locally normalized image patches are convolved with 400 filters and each filter generates a feature map followed by nonlinear activation functions, such as the Rectified Linear Units (ReLU) [18], sigmoid, tanh, etc. In this layer, the kth output feature map y_k can be calculated as follows:

$$y_k = f(w_k * x), \tag{1}$$

where x denotes the input image, w_k stands for the convolutional filter associated with the kth feature map, $*$ indicates the 2D convolution operator, and $f(\cdot)$ is the nonlinear activation function. In this work, ReLU is employed for nonlinear activation due to its efficiency and effectiveness.

2.3. Non-linear Transformation and Normalization

It has been shown in [19] that using a rectifying non-linear transformation operation is an effective way to further improve the CNN performance for visual recognition tasks. This is usually achieved by performing local subtractive or divisive operations for normalization, enforcing a kind of local competition between features at the same spatial location in different feature maps. In this work, the local contrast normalization [19] is carried out with the normalized output y_{kij} produced as

$$y_{kij} = \frac{x_{kij}}{\left(1 + \frac{\alpha}{M_1 \cdot M_2} \sum_{p=i-\frac{M_1}{2}}^{i+\frac{M_1}{2}} \sum_{q=j-\frac{M_2}{2}}^{j+\frac{M_2}{2}} (x_{kpq} - m_{kij})^2\right)^{\beta}}, \tag{2}$$

where the parameters of α and β can be determined using a validation set, which are set to be $\alpha = 0.0001$ and $\beta = 0.75$ empirically in the current work. The local contrast is computed within a local $M_1 \times M_2$ region with the center at (i, j), and m_{kij} is the mean of all x values within the above $M_1 \times M_2$ region in the kth feature map as computed as

$$m_{kij} = \frac{1}{M_1 \cdot M_2} \cdot \sum_{p=i-\frac{M_1}{2}}^{i+\frac{M_1}{2}} \sum_{q=j-\frac{M_2}{2}}^{j+\frac{M_2}{2}} x_{kpq}. \tag{3}$$

2.4. Pooling

The pooling operation is applied on each feature map to reduce the filter responses to a lower dimension. Specifically, each feature map is pooled into one max value and one average value instead of one min value (which is used in [16]).

The max pooling chooses the largest element in each pooling region as

$$y_{kij} = \max_{(p,q)\in\mathcal{R}_{ij}} x_{kpq}, \tag{4}$$

where y_{kij} is the output of the pooling operator related to the kth feature map, x_{kpq} is the element at (p, q) within the pooling region $\mathcal{R}_{ij}$ which represents a local neighborhood around the position (i, j). Regarding the average pooling, it chooses the mean of the elements in each pooling region as

$$y_{kij} = \frac{1}{|\mathcal{R}_{ij}|} \sum_{(p,q)\in\mathcal{R}_{ij}} x_{kpq}, \tag{5}$$

where $|\mathcal{R}_{ij}|$ is the size of the pooling region $\mathcal{R}_{ij}$.

3. EXPERIMENTAL RESULTS

3.1. Database and Evaluation Protocol

In order to demonstrate the performance of the proposed CNN based method for multiply distorted image quality assessment, the LIVEMD [17] database is employed for experiments, which is the most common image database with multiple image distortions. There are two subsets of multiply distorted images in LIVEMD, including 1) blur followed by JPEG compression which is denoted as 'B&J' and 2) blur followed by noise which is termed as 'B&N'. These images are generated by adding different levels of JPEG compression or noises to blurred images, and there are 225 images produced from 15 pristine images in each of these two image subsets. In addition, three experimental scenarios are tested according to which images being used, including 1) 'B&J': only the 'B&J' subset of images are used for training and testing, 2) 'B&N': only the 'B&N' subset of images are utilized for training and testing, and 3) 'ALL': both of the 'B&J' and 'B&N' subsets of images are employed for training and testing. For each experimental scenario, we randomly choose 75% of images for training and the other 25% of images for testing. In order to counterbalance the effect on random data selection, we run the evaluation 3 times to compute the average performance.

As far as performance criteria are concerned, two measurements are adopted to evaluate IQA approaches as suggested in [2], including Spearman Rank-Order Correlation Coefficient (SROCC) and Pearson Linear Correlation Coefficient (PLCC). SROCC is generally used to measure the prediction monotonicity of an IQA metric, which operates only on the rank of data points and ignores the relative distance between data points. Regarding PLCC, it is computed to measure the linear dependence between two quantities after nonlinear regression, which can be performed with the acknowledged logistic mapping function as

$$f(x) = \beta_1 \left(\frac{1}{2} - \frac{1}{1 + e^{\beta_2 (x - \beta_3)}} \right) - \beta_4 x + \beta_5, \tag{6}$$

where $\beta_i, i = 1, 2, \cdots, 5$ are the parameters to be estimated from data.

The proposed CNN based method is compared with a number of IQA metrics, *i.e.*, five representative and outstanding FR-IQA methods including PSNR, SSIM [7], MS-SSIM [8], FSIM [9] and GMSD [10], and four state-of-the-art blind IQA methods, including DIIVINE [12], BLIINDS-II [13], BRISQUE [14] and CNN-KangLe [16]. Moreover, our implementation of the CNN model is derived from the publicly available Caffe toolbox [20].

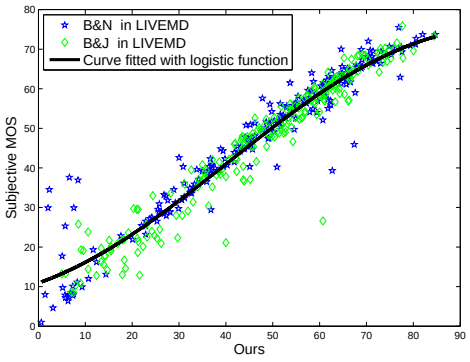

Fig. 2. Scatter distribution and fitted curve achieved by the proposed method on LIVEMD.

Table 1. Comparison of the SROCC performance by different network structures.

Method	B&J	B&N	ALL
CNN-KangLe [16]	0.9627	0.9612	0.9577
CNN-Avg	0.9734	0.9697	0.9669
Ours	0.9853	0.9745	0.9703

Table 2. Comparison of AC by different network structures.

Method	B&J	B&N	ALL
CNN-KangLe [16]	0.4026	0.5541	0.3982
CNN-Avg	0.4486	0.6231	0.4361
Ours	0.4892	0.7088	0.4958

3.2. Comparison with CNN-KangLe [16]

The method CNN-KangLe in [16] has excellent performances for single distorted images, which is not very suitable for multiply distorted images. The main difference between CNN-KangLe [16] and our method mainly includes the pooling strategy and the number of features. In order to clarify the

effectiveness of average pooling and the increment of features employed by our method, we add a model which uses the average pooling instead of the min pooling in [16] and all the other parameters are the same, which is called CNN-Avg in Table 1. As shown in Table 1, the improvement CNN-Avg over CNN-KangLe demonstrates the effectiveness of average pooling. Furthermore, when comparing CNN-Avg and the proposed CNN structure (*i.e.*, 'Ours'), it can be observed that the SROCC performance is further improved by increasing the number of features (*i.e.*, the number of features is increased from 50 in [16] to 400 by our method). Moreover, as suggested in [16], the Accuracy (AC) performances achieved by CNN-KangLe [16] and our proposed method are presented in Table 2, since AC is able to indicate the prediction ability of CNN and a high AC value reveals a high IQA performance in general. From the results shown in Table 2, it is also obvious that the proposed method is better than CNN-KangLe [16] in CNN prediction accuracy for multiply distorted images.

Table 3. Comparison of the SROCC performance under three experimental scenarios ('B&J', 'B&N' and 'ALL').

Method	Type	B&J	B&N	ALL
PSNR	FR	0.6621	0.7088	0.6771
SSIM [7]	FR	0.8493	0.8760	0.8603
MS-SSIM [8]	FR	0.8488	0.8629	0.8363
FSIM [9]	FR	0.8546	0.8644	0.8637
GMSD [10]	FR	0.8247	0.7889	0.8081
DIIVINE [12]	NR	0.7261	0.6120	0.6694
BLIINDS-II [13]	NR	0.6137	0.1074	0.2635
BRISQUE [14]	NR	0.8064	0.2967	0.5342
CNN-KangLe [16]	NR	0.9627	0.9612	0.9577
Ours	NR	**0.9853**	**0.9745**	**0.9703**

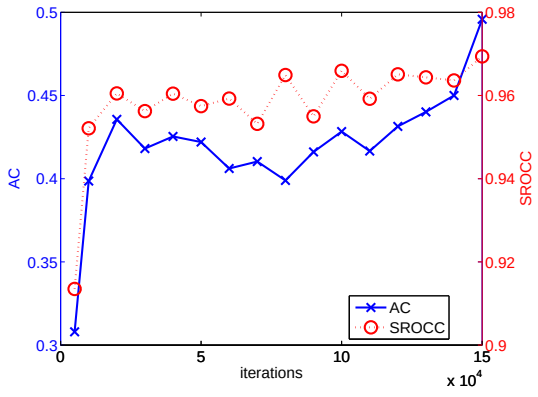

Fig. 3. Trends of SROCC and AC against iterations achieved by the proposed CNN based method.

3.3. Comparison with State-of-the-Arts

The comparative experimental results are presented in Tables 3 and 4 for presenting the SROCC and PLCC performances, respectively. As we know, a better IQA metric is expected to have higher SROCC and PLCC values. From the experimental results, it is observed that our proposed CNN based method is able to achieve the best performance on LIVEMD. Specifically, the performances of 'B&J' is a little better than 'B&N' and the overall performances are also excellent. The scatter distribution and fitted curve of subjective Mean Opinion Score (MOS) of the proposed method are illustrated in Fig. 2, where it can be seen that the obtained scatter distribution correlates consistently with the score points.

Table 4. Comparison of the PLCC performance under three experimental scenarios ('B&J', 'B&N' and 'ALL').

Method	Type	B&J	B&N	ALL
PSNR	FR	0.7425	0.7743	0.7398
SSIM [7]	FR	0.8970	0.8963	0.8915
MS-SSIM [8]	FR	0.8877	0.8914	0.8747
FSIM [9]	FR	0.9065	0.8805	0.8934
GMSD [10]	FR	0.8664	0.8306	0.8462
DIIVINE [12]	NR	0.7983	0.6839	0.7308
BLIINDS-II [13]	NR	0.6309	0.1760	0.3617
BRISQUE [14]	NR	0.8629	0.3683	0.5816
CNN-KangLe [16]	NR	0.9622	0.9547	0.9481
Ours	NR	**0.9858**	**0.9739**	**0.9648**

In the experiments, a total of 150,000 iterations are performed to train the proposed CNN based method. The trends of SROCC and AC performances against the iterations are shown in Fig. 3, where it can be observed that a certain linear relationship between SROCC and AC is obtained.

4. CONCLUSION

In this paper, a CNN based method is proposed to accurately predict multiply distorted image quality without reference images. Based on the proposed method, the relationship between the prediction monotonicity of IQA in terms of SROCC and the prediction accuracy of CNN is revealed. The experimental results also demonstrate that the proposed CNN based method is superior to a number of state-of-the-art IQA approaches for multiply distorted IQA. In the future, we will investigate some fusion techniques which can be applied on CNN, *e.g.*, fusion of support vector machine, restricted boltzmann machine and CNN, to further improve the IQA performances for multiply distorted images.

5. REFERENCES

[1] W. Lin and C.-C. J. Kuo, "Perceptual visual quality metrics: A survey," *J. Visual Commun. Image Rep.*, vol. 22, no. 4, pp. 297–312, May 2011.

[2] "Final report from the video quality experts group on the validation of objective models of video quality assessment," *Video Quality Experts Group (VQEG)*, 2003.

[3] L. Ce, W. T. Freeman, R. Szeliski, and S. B. Kang, "Noise estimation from a single image," in *CVPR'06*, Jun. 2006, pp. 901–908.

[4] G. Zhai, J. Cai, W. Lin, X. Yang, W. Zhang, and M. Etoh, "Cross-dimensional perceptual quality assessment for low bit-rate videos," *IEEE Trans. Multimedia.*, vol. 10, no. 7, pp. 1313–1324, Nov. 2008.

[5] S. Wang, A. Rehman, Z. Wang, S. Ma, and W. Gao, "SSIM-motivated rate-distortion optimization for video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 4, pp. 516–529, Apr. 2012.

[6] A. Rehman and Z. Wang, "Reduced-reference image quality assessment by structural similarity estimation," *IEEE Trans. Image Processing*, vol. 21, no. 8, pp. 3378–3389, Aug. 2012.

[7] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. Image Processing*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[8] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *ACSSC'03*, Nov. 2003, pp. 1398–1402.

[9] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "FSIM: a feature similarity index for image quality assessment," *IEEE Trans. Image Processing*, vol. 20, no. 8, pp. 2378–2386, Aug. 2011.

[10] W. Xue, L. Zhang, X. Mou, and A. C. Bovik, "Gradient magnitude similarity deviation: A highly efficient perceptual image quality index," *IEEE Trans. Image Processing*, vol. 23, no. 2, pp. 684–695, Feb. 2014.

[11] Q. Li and Z. Wang, "Reduced-reference image quality assessment using divisive normalization-based image representation," *IEEE J. Sel. Top. Signal Processing*, vol. 3, no. 2, pp. 202–211, Apr. 2009.

[12] A. Moorthy and A. C. Bovik, "Blind image quality assessment: From natural scene statistics to perceptual quality," *IEEE Trans. Image Processing*, vol. 20, no. 12, pp. 3350–3364, Dec. 2011.

[13] M. A. Saad, A. C. Bovik, and C. Charrier, "Blind image quality assessment: A natural scene statistics approach in the DCT domain," *IEEE Trans. Image Processing*, vol. 21, no. 8, pp. 3339–3352, Aug. 2012.

[14] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Trans. Image Processing*, vol. 21, no. 12, pp. 4695–4708, Dec. 2012.

[15] K. Gu, G. Zhai, M. Liu, X. Yang, W. Zhang, X. Sun, W. Chen, and Y. Zuo, "FISBLIM: A five-step blind metric for quality assessment of multiply distorted images," in *SiPS'13*, Oct. 2013, pp. 241–246.

[16] L. Kang, P. Ye, Y. Li, and D. Doermann, "Convolutional neural networks for no-reference image quality assessment," in *CVPR'14*, Jun. 2014, pp. 1733–1740.

[17] J. Dinesh, A. Mittal, A. Moorthy, and A. C. Bovik, "Objective quality assessment of multiply distorted images," in *ACSS'12*, Nov. 2012, pp. 1693–1697.

[18] V. Nair and G. Hinton, "Rectified linear units improve restricted boltzmann machines," in *ICML'10*, Jun. 2010, pp. 807–814.

[19] K. Jarrett, K. Kavukcuoglu, M. Ranzato, and Y. LeCun, "What is the best multi-stage architecture for object recognition?" in *ICCV'09*, Oct. 2009, pp. 2146–2153.

[20] Y. Jia. (2013) Caffe: An open source convolutional architecture for fast feature embedding. [Online]. Available: http://caffe. berkeleyvision.org/